

# Final Project

## PUBH 8878, Statistical Genetics

### Purpose

- Analyze real genetic data using methods from the course. Choose one of the tracks below (or propose your own) and produce a succinct, reproducible analysis using a public dataset.
- This project emphasizes scientific reasoning, correct statistical practice, clear communication, and reproducible code. It is intentionally open-ended.

### Deliverables

- Project proposal (approx. 1 page): question(s), dataset(s), methods, risks/mitigations, expected outputs. **Due: Wednesday, October 15th by 11:59pm.**
  - Ensure that you are able to access/download your proposed dataset(s)
- Final report (8–12 pages, PDF only): format details below. **Due: Wednesday, November 12th by 11:59pm.**
  - Reproducible materials: a self-contained folder or repo with your `.qmd`/R scripts, environment notes, and seeds. Include a `README` with run instructions.
- 8–10 minute in-class presentation of key results on **Wednesday, November 12th**

### Format

We will follow the format of the journal [Genetics](#). An Overleaf template can be found [here](#). The paper should adhere to the following:

- Begin the paper with an original title, followed by your name, the course, and the date
- The paper should have the following sections:
  - Introduction: state the general problem or issue you are addressing.

- Materials and Methods: describe the methods used to obtain data, analyze data, and to test hypotheses associated with the data.
- Results: describe the results of the data analysis and hypothesis testing.
- Discussion: here you draw conclusions about the problem you studied; this section should include a synthesis of ideas.
- References: List the relevant literature you have read and used to support your arguments/analyze your data. The literature cited should be in the format of the journal Genetics.

## Evaluation (20% of course grade)

- Clarity and scope (20%): well-posed question, appropriate scope for time/resources.
- Methods and correctness (35%): sound statistical modeling, assumptions stated, correct inference, sensible QC.
- Interpretation and communication (25%): figures/tables support claims, limitations, ethical awareness.
- Reproducibility (20%): organized code, seeds, instructions, figures regenerate.

## Example Topics

- Primary GWAS analysis (individual-level data): perform QC, PCA, GWAS, and post-GWAS analyses. Data is typically available for non-human model organisms.
- Secondary GWAS analysis (summary statistics): pick one trait and perform quality checks, Manhattan plot, locus zoom(s), gene/annotation enrichment, and short literature triangulation, SNP-heritability via LD Score Regression, cross-trait genetic correlation.
- Population structure and diversity: PCA/UMAP on a reference genotype panel, compute  $F_{ST}$  between populations, visualize allele frequency spectra, explore LD decay, ADMIXTURE/LEA ancestry components.
- Causal inference with two-sample Mendelian randomization: select a well-powered exposure/outcome with strong instruments, run multiple MR estimators, perform sensitivity and heterogeneity checks, discuss assumptions and violations.
- Fine-mapping or colocalization: focus on 1–2 loci, use LD from a reference panel, apply SuSiE/FINEMAP, test GWAS–eQTL colocalization for a tissue of interest.
- Simulation with real LD: simulate phenotypes on a real genotype panel (e.g., chromosome 22) to study power, inflation, or PRS performance under different architectures.

## Ethics & Responsible Use

- Use population labels with care
- Avoid essentialist interpretations

- Discuss portability and fairness when comparing groups
- Do not attempt re-identification
- Respect each dataset’s license/terms.

## Data Sources (curated)

- [GWAS Catalog \(NHGRI-EBI\)](#): comprehensive registry of GWAS with summary statistics where available; good for trait curation and downloading per-study results.
- [OpenGWAS \(MRC IEU\)](#): programmatic access to >40k GWAS summary-stat datasets; integrates well with R packages `ieugwasr` and `TwoSampleMR`.
- [Pan-UK Biobank \(Broad\)](#): pan-ancestry GWAS results across thousands of phenotypes with interactive PheWeb and bulk download.
- [FinnGen](#): large disease-focused GWAS summary stats and phenotype documentation.
- [Biobank Japan](#): GWAS results across many traits; multi-ancestry comparison opportunities.
- [GIANT Consortium](#): anthropometric trait GWAS (e.g., height, BMI) — classic, clean testbeds.
- [Psychiatric Genomics Consortium \(PGC\)](#): summary stats for psychiatric disorders; read and follow data use terms.
- [1000 Genomes Project \(IGSR\)](#): open, phased whole-genome reference panel with population labels; ideal for PCA, F<sub>ST</sub>, LD, and as an LD reference.
- [HGDP + 1000G combined callset \(gnomAD\)](#): harmonized WGS panel for global structure analyses (VCF/PLINK).
- [gnomAD v4](#): aggregated exome/genome allele frequencies; excellent for frequency-based analyses and QC (not individual-level genotypes).
- [GTEx/eQTL resources](#) and [GTEx v8 summary statistics](#) for colocalization.
- [LD reference \(for LDSC/fine-mapping\)](#) and [baseline annotations](#): precomputed 1000G LD scores.

## Model-organism data sources (genotypes + phenotypes where noted)

- (Mouse) [Mouse Phenome Database \(MPD\)](#): strain, Collaborative Cross (CC), and Diversity Outbred (DO) resources with extensive phenotypes. Many datasets include genotypes or QTL-ready files. Good for GWAS/QTL and replication.
- (Mouse) [International Mouse Phenotyping Consortium \(IMPC\)](#): high-throughput knockout phenotypes with rich metadata. Best for functional interpretation.
- (Mouse) [MGI \(Mouse Genome Informatics\)](#): curated QTL/phenotype annotations and cross references.
- (Rat) [Rat Genome Database \(RGD\)](#): strain genotypes/variants with curated phenotypes, supports QTL/GWAS in rat panels.

- (Drosophila) [FlyBase](#): genome and phenotype annotations, links to population panels and datasets.
- (C. elegans) [WormBase](#): phenotype and functional annotations, pairs well with CeNDR for interpretation.
- (Arabidopsis) [AraGWAS Catalog](#) and [AraPheno](#): GWAS results, phenotypes, and links to [1001 Genomes](#) genotypes, GWAS-ready.

## Helpful Resources by Topic

Here are some resources for topics that we will not cover in depth in this course, if you choose to explore them for this project.

### Causal inference (Mendelian randomization)

- [TwoSampleMR documentation](#) and [ieugwasr](#): end-to-end two-sample MR from OpenGWAS, with instrument selection, Steiger filtering, heterogeneity, and sensitivity analyses.
- [MendelianRandomization \(CRAN\)](#): MR-Egger, IVW, weighted median/mode, useful for triangulation.
- [MR-PRESSO \(CRAN\)](#): outlier detection/correction to assess horizontal pleiotropy.
- [CAUSE](#): robust MR under correlated pleiotropy, helpful when standard assumptions are doubtful.

### Fine-mapping and colocalization

- [susieR](#): Bayesian variable selection and credible sets for fine-mapping; works with in-sample or reference LD.
- [FINEMAP](#): summary-stat fine-mapping with shotgun stochastic search, supports multiple causal variants per locus.
- [coloc](#): Bayesian colocalization testing for two traits (e.g., GWAS–eQTL), simple priors, clear summaries.
- [eCAVIAR](#): probabilistic colocalization allowing multiple causal variants, requires LD and summary stats.

### Simulation with real LD

- [GCTA - simulate phenotypes](#): generate traits on top of real genotype panels (e.g., chr22 PLINK files) using specified architectures.
- [bigsnpr](#): R tooling for large genotype matrices. Simulate phenotypes with real LD and evaluate polygenic methods efficiently.

- [HAPGEN2](#): simulate new genotypes from reference haplotypes (1000G/UKBB) to preserve realistic LD patterns.
- [stdpopsim](#): coalescent simulations with recombination and demographic models, combine with empirical LD panels for hybrid designs.